Factorial designs: an overview with applications to orthodontic clinical trials

Pandis, Nikolaos ; Walsh, Tanya ; Polychronopoulou, Argy ; Katsaros, Christos ; Eliades, Theodore

Abstract: Factorial designs for clinical trials are often encountered in medical, dental, and orthodontic research. Factorial designs assess two or more interventions simultaneously and the main advantage of this design is its efficiency in terms of sample size as more than one intervention may be assessed on the same participants. However, the factorial design is efficient only under the assumption of no interaction (no effect modification) between the treatments under investigation and, therefore, this should be considered at the design stage. Conversely, the factorial study design may also be used for the purpose of detecting an interaction between two interventions if the study is powered accordingly. However, a factorial design powered to detect an interaction has no advantage in terms of the required sample size compared to a multi-arm parallel trial for assessing more than one intervention. It is the purpose of this article to highlight the methodological issues that should be considered when planning, analysing, and reporting the simplest form of this design, which is the 2×2 factorial design. An example from the field of orthodontics using two parameters (bracket type and wire type) on maxillary incisor torque loss will be utilized in order to explain the design requirements, the advantages and disadvantages of this design, and its application in orthodontic research.

DOI: https://doi.org/10.1093/ejo/cjt054

European Journal of Orthodontics 36 (2014) 314–320
doi:10.1093/ejo/cjt054
Advance Access publication 24 July 2013

Factorial designs: an overview with applications to orthodontic clinical trials

Nikolaos Pandis*,**, Tanya Walsh***, Argy Polychronopoulou****, Christos Katsaros* and Theodore Eliades*****

*Department of Orthodontics and Dentofacial Orthopedics, Dental School/Medical Faculty, University of Bern, Switzerland, **Private practice, Corfu, Greece, ***School of Dentistry, University of Manchester, UK, ****Department of Community and Preventive Dentistry, School of Dentistry, University of Athens, Greece, and *****Department of Orthodontics and Paediatric Dentistry, Center of Dental Medicine, University of Zurich, Switzerland.

Correspondence to: Theodore Eliades, Department of Orthodontics and Paediatric Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich 8032, Switzerland. E-mail: theodore.eliades@zzm.uzh.ch

SUMMARY Factorial designs for clinical trials are often encountered in medical, dental, and orthodontic research. Factorial designs assess two or more interventions simultaneously and the main advantage of this design is its efficiency in terms of sample size as more than one intervention may be assessed on the same participants. However, the factorial design is efficient only under the assumption of no interaction (no effect modification) between the treatments under investigation and, therefore, this should be considered at the design stage. Conversely, the factorial study design may also be used for the purpose of detecting an interaction between two interventions if the study is powered accordingly. However, a factorial design powered to detect an interaction has no advantage in terms of the required sample size compared to a multi-arm parallel trial for assessing more than one intervention. It is the purpose of this article to highlight the methodological issues that should be considered when planning, analysing, and reporting the simplest form of this design, which is the 2×2 factorial design. An example from the field of orthodontics using two parameters (bracket type and wire type) on maxillary incisor torque loss will be utilized in order to explain the design requirements, the advantages and disadvantages of this design, and its application in orthodontic research.

Introduction

A common aim of clinical research in dentistry is the evaluation of the effectiveness of different treatment or prevention strategies on clinical or patient-reported outcomes. Among the different clinical research study designs, randomized controlled trials (RCTs) command the highest level in terms of quality in the hierarchy of evidence for the assessment of the effects and safety of an intervention (Moher *et al.*, 2010).

RCTs may be implemented using a plethora of study designs depending on the interventions to be evaluated, the settings, the resources, and practicalities. The most common RCT design explores the effect of two or more interventions at a time in a parallel fashion. Two groups of patients are randomly allocated to the two therapies (or therapy and control) and are followed prospectively. Another design is the crossover (Chan and Altman, 2005), in which, in the simplest form, the same groups of patients are randomly allocated to the two interventions during the first stage, a wash-out period follows, and during the second stage, treatment allocation is reversed so that all patients receive both interventions either during stage 1 or 2. The closest design to crossover in dentistry is the split-mouth design (Pandis *et al.*, 2012). The clustered design (Campbell *et al.*, 2004) allocates interventions to groups of patients and its extension in orthodontics is the design in which multiple observations (teeth nested in patients) are selected per patient (Pandis *et al.*, 2013). The non-inferiority design aims to establish equivalence or non-inferiority of a newer intervention compared with the standard (Piaggio *et al.*, 2006). Finally, the factorial fashion (Montgomery *et al.*, 2003) design is used, in which two or more interventions may be evaluated on the same sample of patients.

The various RCT designs with their different characteristics possess certain advantages and disadvantages, which make them more suitable in specific settings. However, classification is not too rigid as some of the designs may be a hybrid of two or more specific designs (Peters *et al.*, 2003; Bahrami *et al.*, 2004). Additionally, the type of trial design requires different provisions for the number of participants to be included and for appropriate data analysis methodology.

A parallel design randomly assigns one or more interventions to two or more groups of participants, follows

them prospectively, and compares effects between treatment arms. A parallel design may have two or more arms and each participant is randomized to one and only treatment. The parallel design is the most common approach (Chan and Altman, 2005), which however, is not always the most efficient. The loss in efficiency is associated with the fact that when multiple therapies (treatment arms) are investigated, they require many patients in order to get precise estimates, thus increasing trial cost and resources.

In certain situations, it is possible to evaluate two or more interventions simultaneously in a single trial (Hennekens *et al.*, 1996; McAlister *et al.*, 2003; Piantadosi, 2005). This may be accomplished by following a factorial study design. The simplest design takes the form of a 2×2 design (two treatments with two levels each), nevertheless higher-order factorial designs are possible and have been reported (The PARAGON Investigators, 1998; Apfel *et al.*, 2003, 2004; McAlister *et al.*, 2003). The advantages of the factorial design are related to the fact that two or more parameters may be assessed at the same time in the same population simultaneously, thus creating a more efficient trial in terms of resources including sample size compared with separate trials for assessment of each parameter (Montgomery *et al.*, 2003). However, the assumptions that the two treatments may be combined and that there is no interaction (or effect modification) must be satisfied (Ottenbacher, 1991). Interaction (or effect modification) is present when the effect of one variable on an outcome is modified according to the level of a second variable (Altman and Bland, 2003). For example, if we are assessing the effect of the type of orthodontic treatment on maxillary incisor resorption and we find that the effect of the type of appliance is different with different types of wire, then we may say that we have evidence of interaction or effect modification between the intervention (bracket type) and the wire type. In such a situation, a factorial design that would explore the effect of the type of bracket and wire type on root resorption simultaneously in the same sample would not be appropriate. In the case where no interaction exists, a factorial design would probably be an appropriate and efficient method in evaluating the effect of two therapies.

It is the purpose of this article to highlight the methodological issues that should be considered when planning, analysing, and reporting the simplest form of this design, which is the 2×2 factorial design. An example from the field of orthodontics using two parameters (bracket type and wire type) on maxillary incisor torque loss will be utilized in order to explain the design requirements, the advantages and disadvantages of this design, and its application in dental orthodontic research.

Advantages

1. A factorial design is more efficient mainly due to the smaller sample size required (up to one-half) compared with two separate two-arm parallel trials. The efficiency in terms of sample size of the factorial design that tests two interventions at the same time is valid under the assumption that no interaction is present between the two interventions. However, it must be kept in mind that 'interaction tests have low power' and absence of significant interaction is not absolute proof of no interaction (Lubsen and Pocock, 1994).
2. A factorial design is the only design that allows testing for interaction; however, designing a study 'to specifically' test for interaction will require a much larger sample size, and therefore it is essential that the trial is powered to detect an interaction effect (Brookes *et al.*, 2001).
3. Reduced costs, reduced recourses and management needs are found due to the fact that a smaller sample will be required compared with two separate trials.

Disadvantages

1. A factorial design may require extra time, compliance, and management of applying two treatments at the same time.
2. Data analysis and randomization may be a little more complex because participants must be allocated to four arms either in one (A, B, C, and D) or two stages (first intervention and comparator, and then second intervention and its comparator (Montgomery *et al.*, 2003; Machin and Fayers, 2010)
3. Appropriateness and acceptability/tolerance of the combined intervention on biologic and scientific grounds must be explored and determined (Brittain and Wittes, 1989).
4. If interaction is expected, but there is no intention to detect the interaction, the factorial has no sample size advantages compared with two separate two-arm parallel trials. In other words, in this scenario, the sample size will be double the size of the factorial with no interaction or equal to the size of two 2-arm parallel trials (Brookes *et al.*, 2001). Therefore, the investigators must be as sure as possible that no interaction is present between the two interventions when undertaking a factorial design in which they would like to assess simultaneously the effect of two treatments. If the objective is to specifically detect interaction, the sample increases by 4-fold compared with the factorial with no interaction when we want to observe an interaction effect equal to the effect to be detected between the two arms of the parallel trial (Brittain and Wittes, 1989; Brookes *et al.*, 2001; Montgomery *et al.*, 2003; Piantadosi, 2005)

Randomization

Randomization in factorial designs may follow similar and appropriate methods used with parallel trials, such as

simple, restricted, stratified randomization, or minimization (Pandis *et al.*, 2011). One difference is that individuals must be randomized more than once depending on the factorial design. In a 2 × 2 factorial design, participants may be randomized to either the experimental or the control group for intervention A and then to either experimental or control group for intervention B. Alternatively, they may be randomized simultaneously in the four groups of the 2 × 2 factorial design (Montgomery *et al.*, 2003; Machin and Fayers, 2010).

Sample size for a 2 × 2 factorial design

When the main reason for the trial is to compare the separate impacts of two interventions within the same trial, the approach to sample size calculations is relatively straightforward and it is common to consider the trial as two separate two-arm trials. The sample size for each of the separate comparisons is calculated and whichever of these results in the largest number of patients provides the basis for the overall sample size. These separate calculations are likely to be similar if the same outcome is used for both (Brookes *et al.*, 2001; Montgomery *et al.*, 2003). However, if the outcome and/or the assumptions are 'different', then the required sample for each intervention may be different. When the study must be powered to specifically detect interaction, the factorial design loses its efficiency as the required sample size must be increased dramatically. For example, to be able to detect an interaction effect equal to the effect of the treatments under study, a 4-fold increase in the sample is required (Brookes *et al.*, 2001; Montgomery *et al.*, 2003).

Example of factorial design in orthodontic research

We are interested in evaluating the amount of torque loss/final position of maxillary incisors during retraction in first maxillary premolar extraction class II/1 cases. We would like to compare the effect on torque loss/maxillary incisor position for both wire types 0.019 × 0.025 stainless steel (SS) and 0.19 × 0.25 reverse curve NiTi (RC-NiTi) and bracket type [Self-ligating (SLB) or Conventional (CB)] at the same time. The details for this 2 × 2 factorial design are shown in the upper part of Table 1.

One way to analyse the data from this trial would be to perform pair-wise comparisons among all available groups shown in Table 1 (lower part). This analysis will compare A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D. This approach, although often used, has the following problems.

It invokes the problem of multiplicity and the likelihood of false-positive findings related with multiple comparisons. It is expected that at an alpha level of 5 per cent, for every 20 tests, one test shall be positive only by chance. In other words, as we increase the number of statistical comparisons, the probability of observing a statistically significant finding just by chance increases. This has been well documented in the biomedical literature (Oxman and Guyatt, 1992; Assmann *et al.*, 2000).

Table 1 Factorial design for simultaneously assessing the effect of wire type and bracket type on torque loss during maxillary anterior teeth retraction in class II/1 extraction cases.

		Wire type	
		0.19 × 0.25 Stainless steel	0.19 × 0.25 Reverse curve NiTi
Bracket type	Self-ligating	A	B
	Conventional	C	D
Subgroup comparisons		A versus B	
		A versus C	
		A versus D	
		B versus C	
		B versus D	
		C versus D	
Main effects comparisons		A+B versus C+D	
		A +C versus B + D	

Possible comparisons among subgroups and for main effects (lower part of the table).

False-positive results may lead to over-interpretation of findings based solely on *P* values, selective reporting, and publication bias (Hahn *et al.*, 2000). Investigators may be tempted to focus, in the presentation of their results, on what is statistically significant and not on what is clinically significant. Readers may interpret research findings on the basis of statistical significance or no significance, with little regard to clinical importance, as there is a misconception that a low *P* value means a strong clinical effect (Goodman, 1999). Additionally, it has been reported in biomedical literature in general, and also in orthodontics, that studies with statistically significant findings are more likely to be published compared with studies reporting no significant findings (Moscati *et al.*, 1994, Rosenthal, 1979, Scholey and Harrison, 2005; Hopewell *et al.*, 2007; Koletsi *et al.*, 2009). Therefore, if only a subsample of the trials is published, then clinical decisions may be based on only a part of the existing evidence.

Because comparisons are performed on subgroups, these tests have low power as the subgroups have smaller samples in relation to the calculated sample. Low power may hide a clinically important effect if conclusions are based only on *P* values (Yusuf *et al.*, 1991). This statement may contradict the previous point; however, during subgroup analyses, power is lost; additionally a strong effect may appear, which could be a chance finding.

The above approach that resorts to subgroup comparisons defeats the purpose of a factorial design as the selected comparisons require larger sample sizes.

A more appropriate approach to data analysis would be to make the following comparisons under the assumption

of no interaction between wire type and bracket type; then we can conduct the two following comparisons of the main effects (Table 1, lower part).

1. A+B versus C+D. This compares the torque loss between the two bracket types and assumes that torque loss difference between SLB versus CB does not change as the type of wire changes (no interaction assumption). In other words, torque loss difference is the same for, let us say, SLB versus CB regardless of using the regular SS wire or the reverse curve wire.
2. A+C versus B+D. This compares the torque loss between the wire types and assumes that torque loss difference between SS versus RC-NiTi does not change as the type of bracket changes (no interaction assumption). In other words, torque loss difference is the same for, let us say, the SS versus RC-NiTi wire regardless of the use of SLB or CB.

We can determine the required sample size by calculating the sample size for both comparisons (A+B versus C+D and A+C versus B+D) and then take the larger sample size required (if different). So, let us say that the expected torque loss in the SS wire is 10 degrees and that we would like to be able to observe a 3-degree difference between the two wire types, with alpha = 0.05 and power = 90 per cent. We assume the standard deviations (SD1, SD2) for both groups are the same and equal to 5 degrees. Sample calculations are based on assumptions that are derived from previous publications or from piloting. The assumption of equal standard deviations is common, but it could be easily changed and applied according to the specific circumstances.

We will use the following formula (Pocock, 1983):

$$n = f(\alpha,\beta)\,\chi\,\frac{2\mathrm{SD}^2}{(\mu 1-\mu 2)^2}.$$

Formula 1. This formula is used for sample calculation for two means, normally distributed quantitative outcomes, equal trial-arm allocation ratio, and two-sided tests, where $\mu 1$ = anticipated mean torque loss on the standard treatment (CB), $\mu 2$ = anticipated mean torque loss on the alternative treatment (SLB), SD = standard deviation for torque loss (assumed the same on both arms), α = type I error (significance level), β = type II error (1 − β = power), and $f(\alpha,\beta)$ is a function of α and β derived from the standard normal distribution and their values are given in Table 2.

$$n = f(\alpha,\beta)\,\chi\,\frac{2\mathrm{SD}^2}{(\mu 1-\mu 2)^2} = 10.5 \times \frac{2\times 5^2}{3^2} = 58.33\ .$$

Therefore, the answer is 58.33 per treatment arm for a total of 118 patients (rounded up), and this is the sample size for the comparison of treatment arms A+C versus B+D.

Table 2 Values of the function $f(\alpha,\beta)$ for different values of alpha and beta.

		β			
		0.05 (95% power)	0.1 (90% power)	0.2 (80% power)	0.5 (50% power)
α	0.05	13.0	10.5	7.85	3.84
	0.01	17.8	14.9	11.7	6.63

With the same assumptions for the comparison between bracket types A+B versus C+D, the sample calculation will yield again the same number per treatment arm (59).

Therefore, in the absence of interaction, we can assess the effects of both bracket type and wire type at the same time using the sample required to assess the effect of only one comparison, which in this example will be 59 patients per arm or a total of 118 patients. Provisions for losses to follow-up should also be considered. If the assumptions were different in terms of the expected mean values and variances for one of the main effects comparison, then a different sample size would have resulted from the calculation. In this scenario, the larger sample from the two calculations would have been required.

In the presence of interaction, the factorial design requires a sample size similar to the size required for two separate two-arm parallel trials (four-arm trial) and therefore there is no real advantage in terms of sample size (Brookes *et al.*, 2001; Montgomery *et al.*, 2003; Wang and Bakhai, 2006). In this case, comparisons should be performed within strata and if no upward sample adjustments are made, the study would be underpowered. Therefore, when we expect an interaction and the primary intention of the study is not to detect the interaction, the 2 × 2 factorial designs becomes a four-arm trial and sample sizes are determined accordingly. In this scenario, two separate trials could be carried out, one comparing torque loss for SLB versus CB and one comparing torque loss for SS versus RC-NiTi. A factorial design powered to detect interaction is a very useful tool, if not the only one available, to assess whether the effect of one parameter depends on the other parameter under investigation (Wang and Bakhai, 2006).

Finally, when designing a trial to 'specifically' detect a level of interaction (3 degrees in this example) equal to the difference to be detected between the two treatment arms (either in wire type or bracket type), the required sample size must be increased four times compared with the same design with no interaction for a total of 472 participants (Brookes *et al.*, 2001; Montgomery *et al.*, 2003).

We will proceed with sample calculation for interaction in more detail. We showed earlier that if we want to detect a difference of 3 degrees between bracket types or wire types (same assumptions for both interventions), SD1 = SD2 = 5

degrees, power = 0.90, and alpha = 0.05; in the two-arm parallel-trial scenario, we would need a total 118 participants for both arms.

Suppose now that we want to conduct a factorial design trial for wire type and bracket type on torque loss with the objective to specifically assess interaction.

The main effects and the interaction comparisons will be the following.

1. Main effect for wire: the treatment effect of SS versus RC-NiTi wire regardless of bracket type.
2. Main effect for bracket: the treatment effect of SLB versus CB regardless of wire type.
3. Interaction: Torque loss (SLB/SS − SLB/RC-NiTi) − Torque loss (CB/SS − CB/RC-NiTi) or Torque loss (SS/SLB − SS/CB) − Torque loss (RC-NiTi/SLB − RC-NiTi/CB)

The next step, as in the usual sample size calculations, would be to decide what would be the minimum difference of clinical importance that we would like to detect. Previously, we used as a minimum difference 3 degrees for the main effects comparison (scenarios 1 and 2 above) and we will use the same difference for the interaction comparison (scenario 3 above; Altman & Bland, 2003). We assume the standard deviation is equal in all four subgroups (SD1 = SD2 = SD3 = SD4) and that it is 5 degrees. When the objective of the study is to specifically detect interaction, the required sample size must be increased dramatically (4-fold in this example; Brookes *et al.*, 2001).

When we compare two groups, the standard deviation used for the test is not SD1 minus SD2 but SD1+ SD2, because the standard deviation of the difference of the comparison groups is expected to be higher than the individual group standard deviations. This is counterintuitive because we would think the standard deviation of the difference is equal to the difference of the standard deviations, and it is important to realize that it is the sum of the standard deviations. From Formula 1, we know that as the standard deviation increases, so does the required sample size. For the interaction test, we have four subgroups and to assess their combined difference, we calculate the standard deviations as the sum of all subgroup standard deviations [= 4 SD], which would give a variance equal to $4 \times 5^2 = 100$, which is the variance for the interaction test including four subgroups (each subgroup variance = $SD^2 = 25$).

Statistical analysis

The statistical analysis selected will depend on the type of outcome and the research question. A classic approach for the 2 × 2 factorial designs when the outcome is continuous as in our example (torque loss in degrees) is the two-way analysis of variance (two-way ANOVA), similar to a multivariable linear model with two predictors. The regression model may be written as follows:

$$y = \beta_0 + \beta_{\text{bracket}} \times \chi_{\text{bracket}} + \beta_{\text{wire}} \times \chi_{\text{bracket}} \quad (1)$$

Here, y is the outcome measurement of torque loss in degrees, β_0= the expected torque loss in degrees for the reference bracket (CB) and wire (SS) groups, $\chi_{bracket}$ = 1 and 0 for bracket SLB and bracket CB, respectively, and χ_{wire} = 1 if RC-NiTi wire is given and 0 for SS wire. In this equation, we selected CB and SS as the baseline or reference groups, but we could have easily selected SLB and RC-NiTi as the reference and modified the interpretation accordingly.

To test for interaction between bracket and wire, Equation 1 may be expanded as follows:

$$y = \beta_0 + \beta_{\text{bracket}} \times \chi_{\text{bracket}} + \beta_{\text{wire}} \times \chi_{\text{bracket}} + \beta_{\text{bracket}_{\text{wire}}} \times \chi_{\text{bracket}_{\text{wire}}} \quad (2)$$

Here, y is the outcome measurement (torque loss) in degrees; β_0, χ_{bracket}, χ_{wire} are the same as for Equation 1, and $\beta_{\text{bracket}_{\text{wire}}} \times \chi_{\text{bracket}_{\text{wire}}}$ is the interaction term. The interaction term may be considered as the value that the estimates should be adjusted for in order to get the correct values when we assume that the effect of bracket type is influenced by the effect of wire type.

In the current example, the main analysis computes only main effects, i.e. the effects of bracket type and wire type on torque loss independently as there is no interaction assumption. Depending on the type of the intervention, it is natural to be interested to know whether the effect of treatment may be different between subgroups. As explained earlier, subgroup analyses have certain problems associated with them. A better approach to test for possible differences would be to perform an interaction test as shown in Equation 2 (Yusuf *et al.*, 1991; Assmann *et al.*, 2000) However, interaction tests have low power and if the objective is to test for the presence of interaction or to compare certain subgroups, the study should be powered accordingly as it is incorrect to select a sample based on a certain comparison and then use the same sample to make comparisons not intended during the pre-trial sample calculations.

Informal assessment of interaction

We can conduct a statistical test to assess the presence of interaction; however, as already mentioned, these tests suffer from low power. We can also conduct an informal interaction test by looking at the tabulated results under two scenarios of torque loss differences (Table 3). Again, we have interaction when the effect of bracket type on torque loss measured in degrees is different at the two different levels of the variable 'wire type'. Therefore, we would like to see whether torque loss difference between SLB versus CB is the same for patients with SS wire and those with RC-NiTi wire and we can study this as follows.

If there is no interaction, the difference in torque loss between CB and SLB should be similar in both SS and RC-NiTi wire patients, and if there is interaction, the difference in torque loss between the bracket CB and SLB should be different between SS and RC-NiTi wires. In Table 3, the differences in torque loss (between CB and SLB, subpart a) are similar (1 versus 2 degrees) regardless of the type of wire; in this case, no interaction is suspected. On the contrary, in section 'b' of Table 3, the differences in torque loss (between CB and SLB) are large (3 versus 10 degrees), indicating presence of interaction (Matthews and Altman, 1996a,b).

The same question could be asked the other way around. Is the difference in torque loss between SS and RC-NiTi groups modified depending on the type of bracket? Comparing the differences by row or by column is a quick method for checking for interaction without statistical testing. However, it should be kept in mind that the presence or absence of interaction may depend on the scale of measurement. For example, absence of interaction on an additive scale may not preclude absence of interaction on a multiplicative scale (Brittain and Wittes, 1989).

To further elaborate on the issue of subgroup comparisons versus interaction testing, it is likely that if we adopt subgroup comparisons like SLB versus CB separately within the SS and RC-NiTi groups and the sample size is different between subgroups, it is possible to obtain conflicting results. For example, as the *P* value depends on sample size and variance, even though the clinical difference is small and indicates no interaction, the *P* value may be significant in one of the subgroup comparisons (Table 4). This type of problem is avoided with the use of an interaction test and conclusions are not drawn based on *P* values from underpowered subgroup analyses (Altman and Bland, 2003).

Table 3 Tabulation for informal assessment of interaction.

a.	**0.019×0.025 Stainless steel**	**0.019×0.0 25 NiTi**	**Difference**
Self-ligating	5	4	1
Conventional	7	5	2
Difference	2	1	—
b.	**19×25 SS**	**19×25 NiTi**	**Difference**
Self-ligating	5	2	3
Conventional	13	3	10
Difference	8	1	—

a. shows that the difference in effect between self-ligating and conventional brackets is similar in the presence of either the 0.19×0.25 SS wire (2 degrees) or the presence of the 0.019×0.025 NiTi wire (1 degrees). Similarly, the difference between wire types is similar in the presence (1 degree) or absence of the self-ligating appliance (2 degrees)
b. shows that the difference in effect between self-ligating and conventional brackets is different in the presence of either the 0.19×0.25 SS wire (8 degrees) or the presence of the 0.019×0.025 NiTi wire (1 degrees). Similarly, the difference between wire types is similar in the presence (3 degree) or absence of the self-ligating appliance (10 degrees)

Table 4 Subgroup comparisons may yield conflicting results if the focus is on statistical significance as *P* values depend on sample size and variance.

	0.019×0.025 Stainless steel	0.019×0.0 25 RC-NiTi	Difference	*P* value
Self-ligating	5	4	1	<0.05
Conventional	7	5	2	>0.05
—	2	1	—	—

In the first test, we are assuming large sample size, and in the second, a small sample size, whereas standard deviation is assumed the same for all group means.

Reporting

Reporting of factorial designs should follow the guidelines proposed by the Consolidated Standards of Reporting Trials (CONSORT) statement as closely as possible (Moher *et al.*, 2010); however, specific guidelines for factorial designs are not yet available. A key issue is that in case interaction is detected, then estimates should be reported per stratum or estimates should be calculated after considering the calculated value of the interaction term (Lubsen and Pocock, 1994). In other words, if together with the main effects the interaction term is calculated after applying a regression model, the correct estimates that incorporate the interaction effect can be easily calculated. If there is interaction that cannot be detected due to low power when sample size for the factorial design is selected under the no interaction assumption, then the problem of interpretation will depend on whether the interaction is qualitative or quantitative. For quantitative interaction, usually the issue would be that the main effects will overestimate the effects for some individuals and underestimate them for some others. Although the interaction and the means of the four cells must be presented, the main effects may still be a reasonable representation of the intervention effects either separately or combined. On the other hand, if the interaction is qualitative, such change of direction of effect between subgroups and presenting the combined results would be most likely misleading (Montgomery *et al.*, 2003).

Conclusion

- A factorial design of an RCT allows assessment of two treatments at the same time on the same sample.
- If the conditions are satisfied (no interaction between the two treatments, interventions may be combined), the factorial design allows using half of the sample required for the corresponding two separate two-arm parallel trials.

- The factorial design is the only approach that allows the assessment of two or more interventions simultaneously and the evaluation of interactions. If the objective of the factorial design is to detect interaction(s), the sample size must be dramatically increased.

References

Altman D G, Bland J M 2003 Interaction revisited: the difference between two estimates. British Medical Journal 326: 219

Apfel C C *et al.* 2003 An international multicenter protocol to assess the single and combined benefits of antiemetic interventions in a controlled clinical trial of a 2x2x2x2x2x2 factorial design (IMPACT). Controlled Clinical Trials 24: 736–751

Apfel C C *et al.*; IMPACT Investigators 2004 A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. The New England Journal of Medicine 350: 2441–2451

Assmann S F, Pocock S J, Enos L E, Kasten L E 2000 Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355: 1064–1069

Bahrami M *et al.* 2004 Effectiveness of strategies to disseminate and implement clinical guidelines for the management of impacted and unerupted third molars in primary dental care, a cluster randomised controlled trial. British Dental Journal 197: 691–696

Brittain E, Wittes J 1989 Factorial designs in clinical trials: the effects of non-compliance and subadditivity. Statistics in Medicine 8: 161–171

Brookes S T, Whitley E, Peters T J, Mulheran P A, Egger M, Davey Smith G 2001 Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment 5: 1–56

Campbell M K, Elbourne D R, Altman D G; CONSORT Group 2004 CONSORT statement: extension to cluster randomised trials. British Medical Journal 328: 702–708

Chan A W, Altman D G 2005 Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 365: 1159–1162

Goodman S N 1999 Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine 130: 995–1004

Hahn S, Williamson P R, Hutton J L, Garner P, Flynn E V 2000 Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. Statistics in Medicine 19: 3325–3336

Hennekens C H *et al.* 1996 Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease New England Journal of Medicine 334: 1145–1149

Hopewell S, Clarke M, Stewart L, Tierney J 2007 Time to publication for results of clinical trials. Cochrane Database Syst Rev 2: MR000011.

Koletsi D, Karagianni A, Pandis N, Makou M, Polychronopoulou A, Eliades T 2009 Are studies reporting significant results more likely to be published? American Journal of Orthodontics and Dentofacial Orthopedics 136: 632.e1–632.e5

Lubsen J, Pocock S J 1994 Factorial trials in cardiology: pros and cons. European Heart Journal 15: 585–588

Machin D, Fayers P M 2010 Randomized Clinical Trials: Design, Practice and Reporting. Wiley-Blackwell, Chichester

Matthews J N, Altman D G 1996a Interaction 2: compare effect sizes not P values. British Medical Journal 313: 808

Matthews J N, Altman D G 1996b Interaction 3: how to examine heterogeneity. British Medical Journal 313: 862

McAlister F A, Straus S E, Sackett D L, Altman D G 2003 Analysis and reporting of factorial trials: a systematic review. Journal of the American Medical Association 289: 2545–2553

Moher D *et al.* 2010 CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. British Medical Journal 340: c869

Montgomery A A, Peters T J, Little P 2003 Design, analysis and presentation of factorial randomised controlled trials. Biomed Central Medical Research Methodology 3: 26

Moscati R, Jehle D, Ellis D, Fiorello A, Landi M 1994 Positive-outcome bias: comparison of emergency medicine and general medicine literatures. Academic Emergency Medicine 1: 267–271.

Ottenbacher K J 1991 Interpretation of interaction in factorial analysis of variance design. Statistics in Medicine 10: 1565–1571

Oxman A D, Guyatt G H 1992 A consumer's guide to subgroup analyses. Annals of Internal Medicine 116: 78–84

Pandis N, Polychronopoulou A, Eliades T 2011 Randomization in clinical trials in orthodontics: its significance in research design and methods to achieve it. European Journal of Orthodontics 33: 684–690

Pandis N, Walsh T, Polychronopoulou A, Katsaros C, Eliades T 2012 Cluster randomized clinical trials in orthodontics: design, analysis and reporting issues. European Journal of Orthodontics 35: 669–675

Pandis N, Walsh T, Polychronopoulou A, Katsaros C, Eliades T 2013 Split-mouth designs in orthodontics: an overview with applications to orthodontic clinical trials. European Journal of Orthodontics 35: 783–789

Peters T J, Richards S H, Bankhead C R, Ades A E, Sterne J A 2003 Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. International Journal of Epidemiology 32: 840–846

Piaggio G, Elbourne D R, Altman D G, Pocock S J, Evans S J; CONSORT Group 2006 Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. Journal of the American Medical Association 295: 1152–1160

Piantadosi S 2005 Clinical trials: a Methodologic Perspective. 2nd edn John Wiley, New York

Pocock S J 1983 Clinical Trials: a Practical Approach. Wiley, Chichester

Rosenthal R 1979 The file drawer problem and tolerance for null results. Psychological Bulletin 86: 638–641

Scholey J M, Harrison J E 2005 Delay and failure to publish dental research. Evidence-Based Dentistry 6: 58–61

The PARAGON Investigators 1998 International, randomized, controlled trial of lamifiban (a platelet glycoprotein IIb/IIIa inhibitor), heparin, or both in unstable angina. Circulation 97: 2386–2395

Wang D, Bakhai A 2006 Clinical trials in practice. A practical guide to design, analysis and reporting, chapter 10. Remedica, London

Yusuf S, Wittes J, Probstfield J, Tyroler H A 1991 Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. Journal of the American Medical Association 266: 93–98